DOCUMENT RESUME

ED 164 611 TM 008 180

AUTHOR Popp, Jerome A.
TITLE Toward a Quantitative Estimate of Internal Validity.
PUB DATE 28 Mar 78
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Data Collection; Generalization; Hypothesis Testing; Mathematical Models; *Predictor Variables; *Research Design; Research Methodology; Scientific Methodology; *Validity
IDENTIFIERS *Internal Validity

ABSTRACT
Internal validity is described as a matter of how well a particular instance of data collection or generation can be described and explained. It is a property of the procedures used in the collection or generation of data. The notion of internal validity is examined in order to establish a method of quantitatively estimating it. A coefficient of internal validity is defined by the equation that it equals one divided by one more than the number of alternative independent plausible hypotheses. If experimented procedures rule out all alternative plausible hypotheses, then the internal validity coefficient equals one. As the number of alternative hypotheses approaches infinity, the internal validity coefficient approaches zero. Alternative hypotheses may be considered to be equivalent to alternative independent variables that are discrete and non-overlapping and cannot be ruled out and are independently definable. A suggestion is given for weighting independent variables according to their rank order of being plausible. This view is coherent with Harman's view of induction as inference to the best explanation. (Author/CTM)

ED164611

TOWARD A QUANTITATIVE ESTIMATE
OF
INTERNAL VALIDITY

Jerome A. Popp
Chairman, Department of Foundations
of Education
Southern Illinois University
Edwardsville, IL 62026

Presented to Division D,
of the American Educational Research Association,
March 28, 1978, Toronto, Canada

ABSTRACT

The notion of internal validity is examined for purposes of establishing a method of quantitatively estimating it. The logical problem of identifying independent alternative hypotheses is considered as well as their relative plausibility weighting. Finally, the question of internal validity is viewed as being a form of Harmon's "inference to the best explanation".

TOWARD A QUANTITATIVE ESTIMATE
OF
INTERNAL VALIDITY

Analyses of the conclusions about how the results of experiments should be interpreted are commonly conducted in terms of external and internal validity. The continued elaboration of these ideas would, it seems, increase the adequacy of these analyses. The present paper explores the nature of the latter type of validity in an attempt to further refine this concept within educational research. This attempt seems to involve (1) rendering more precise our intuitions or our working notions about internal validity, and (2) marking the kinds of logical problems encountered in trying to give such an account.

Studies are often ordered in terms of internal validity. One study is judged as possessing greater internal validity than another, indicating that the concept is basically a quantitative notion. The present paper attempts to develop a method for the quantitative estimation of internal validity. While there are pitfalls involved in such an effort, it is hoped that the effort will stimulate an interest in the problem by logic-minded educationists.

I. The Problem of Internal Validity

To keep the discussion as concrete as possible within the demands of purpose, consider the notorious one group pretest/post-test design. Assuming that the pre- and post-readings on the dependent

variable(s) reveal a statistically significant gain on the post-scores, the question becomes: what produced this gain; or, how can the observed change in the dependent variable best be explained?

One readily available candidate, and one in which there will be great interest, is of course the independent variable(s), i.e., the treatment(s) explains the observed difference. Or more accurately, the hypothesis asserting a relation between the independent variable and the dependent variable can serve in the explanation of the observed difference. There are, however, alternative explanations of this gain. The weaker the design of the study the more alternatives one can find. Intuitively, it seems that the more of these alternatives which are plausible, the lower the internal validity of the study. The question of internal validity is taken to be: Did the treatment make any difference or have any effect upon the criterion variable? If a change in the criterion readings is observed, how sound or warranted is the explanation of this difference by the research hypothesis? If no change is observed, then did the treatment have an effect which was masked by other extraneous factors? Internal validity is a matter of how well a particular instance of data collection or generation can be described and explained.

From these basic notions, it is seen that the property <u>internal validity</u> is a property of the procedures used in the collection or generation of data on a particular occasion. Internal validity is a variable for the population of data gathering methods. Moreover, it is not taken to be a categorical property for the logic-

in-use or at least the language-in-use refers to "the degree of internal validity"; thus it seems that a reasonable approach to the concept of internal validity is to view it as a quantitative variable.

Internal validity is usually mentioned in the context of experimental research. There is talk of the two kinds of experimental validity. It seems that one could properly speak of the internal validity of <u>ex post facto</u> studies. One can in these cases ask, did the hypothesized independent variable cause the observed changes in the dependent variable? Manipulation of the independent variable allows one to know more about the independent variable than do most <u>ex post facto</u> studies; but this amounts to one's having a higher degree of confidence in the description of the variance of the independent variable than most <u>ex post facto</u> studies allow. It does not, however, show that the question of internal validity, which is a question of how data is to be interpreted, is not appropriate to <u>causal comparative</u> work. It seems that the question of internal validity is relevant to any methods used to gather data relevant to any hypothetical causal-relationship.

II. Internal Validity Measure Function

The problem of internal validity is one of how well certain particular events can be explained. As noted above, there will be much interest at the conclusion of data collection in determining how well the research hypothesis can explain the results obtained. To test the ground for the treatment variable explanation, the following procedure is suggested.

(1a) Assume the data reports are correct and that there was a change in the dependent variable. ((1b) would be the no change case and will not be considered here though the following applies to it as well.) If the assumption that the data reports are correct cannot be made for whatever reason, then the question of internal validity evaporates. There is no question of how to explain a single event if the nature of that event is unknown.

(2) Assume that the treatment variable did not produce the change in the dependent variable.

(3) Ask: what produced change? Or, how can this change be explained?

(4a) If upon careful examination of the research procedures no reasonable or plausible explanations can be found which are consistent with (1a) and (2), then one is forced by rationality to reject either (1a) or (2). If the truth of (1a) is not established then this procedure is unnecessary. The rejection of (2) is the acceptance of the research hypotheses, i.e., of the treatment-variable explanation. This conclusion is epistemically open and practically closed. It is possible that one may detect or construct an explanation of the results at some future time. One cannot know for sure that no answer to (3) is possible. Conclusion (4a) is open in this sense.

With respect to the practice of science, however, one must view things differently. In science there is always the possibility of some further data forcing the reinterpretation of past findings producing new conclusions. The scientist must, nevertheless, show the best conclusions he can on the basis of the total evidence available.

(4b) If one could find explanations of the observed changes which are consistent with (1a) and (2), then the veracity of the treatment explanation is suspect. There exist competing explanations of what happened; and there is no way to choose between these competing hypotheses within the limits of the data of this particular study. Internal validity should be a function of the number and quality of these alternative hypotheses. The greater the number of such, the lower the degree of internal validity. In cases where no alternative can be found, (4a), internal validity should be highest, decreasing as the list of alternatives grows.

Consider the following equation, where V_i is the internal validity and N is the number of alternative hypotheses:

$$V_i = \frac{1}{N + 1}$$

This definition of internal validity meets the requirements of our intuitions. It assigns the internal validity unity when there are

no alternative hypotheses. As the number of alternatives approaches infinity, the function value approaches zero.

Several conceptual questions now present themselves. It is obviously critical how one counts alternative hypotheses. What is required is a way of determining the logical independence of the proposed alternative hypotheses. Secondly, some may question the above function on the grounds that it treats all alternative hypotheses as equally well supported; but is it not the case that certain alternative hypotheses are more plausible than others?

III. Independent Alternative Hypotheses

Let S be the set of non-refuted alternative hypotheses for a given design. Assume that this set is practically complete or closed (see above section). But what does it mean to say that any given hypothesis is refuted? It of course is not intended to suggest that science is to produce absolute knowledge. To be relevant a hypothesis must present an independent variable which is linked hypothetically to the dependent variables whose measures constitute the data; the hypothesis must explain the data at hand. A refuted hypothesis which explains the data is a hypothesis which is ruled out on one of three grounds:

(1) Other well-accepted hypotheses contradict this explanation.

(2) This design used to produce these data rules out this explanation of the data.

(3) Our metaphysical assumptions render this explanation "impossible".

(1) When reviewing the data, one will resist using a hypothesis to explain or interpret what happened which contradicts other well-supported relationships. It is not that we never question that which is "established" but only that we will make such challenges only after we have established the internal validity of our design. All research studies hang together as it were. (2) One of the major purposes of the creation of a research design is to rule out alternative after-the-fact accounts of our results. We plan ahead with regard to being able to draw defensible conclusions. (3) One could construct alternative hypotheses as to why the experimental group did better than the control group, by referring to entities such as demons--there are invisible demons who like program learning and who always confuse people who do not use this method, like our control group who was not taught by the program method. Such explanations are too farfetched; but to realize that scientific inquiries do operate out of a basic metaphysical framework, or "blueprint" as Maxwell calls it, is a very important part of our conception of the nature of science. See Maxwell (1974).

Given that we have identified the set S (all non-refuted alternative hypotheses), we encounter the question of the uniqueness of this set. Can this list be given in such a fashion that when it is counted a stable number can be obtained? Or stated differently, is the membership of S uniquely describable? What we require is a method of writing a basic list of the members of S.

Each member of S will have the same dependent variable(s); thus, what we are actually asking for is a list of alternative independent variables—alternatives to the treatment variable(s), t. We want a list of discrete, non-overlapping independent variables which have not been ruled out. Just as we require the independent and dependent variables of any hypothesis to be independently definable—if they are not, the hypothesis becomes true (or false) or partially true (or false) by definition—we require that the independent variables of the members of S also be independently definable. In other words, the definiens of each independent variable(s) of the members of S must be mutually distinct. This guarantees that we are dealing with discrete alternatives. Moreover, it rules out the possibility that one independent variable of the members of S will entail another. The independent variables associated with S will thus be logically basic or atomic, as it were.

This appeal to logically discrete definiens, brief as it is, does resolve the first problem: how to determine the logical independence of the elements of the set of alternative hypotheses to the research hypothesis.

IV. Plausibility of Alternative Hypotheses

The second problem mentioned above was in effect a rejection of our definition "V_1". That function treats all alternative hypotheses as if they were equally meritorious; but some alternative independent variables are going to be more plausible than others. In some

experimental situations, testing will be a better bet than maturation, yet this seems to be ignored by the simple-minded function given above.

It would be an improvement to rank alternative hypotheses or independent variables including the treatment variable (5) in this list at the appropriate rank. This list will have N + 1 members since N is the number of alternative hypothesis with respect to t. Assign the weight of N + 1 to the first member of the list, (N + 1) - 1 to the second, and so forth. Sum the weights. Internal validity can be defined as follows:

$$V_i = \frac{\text{Weight of } \underline{t}}{\text{Sum of weights}}$$

Five situations are presented below, together with their generated internal validity measures, for illustrative purposes.

Case	1	2	3	4	5
Rankings	t	t	t	A_1	A_1
		A_1	A_1	t	A_2
			A_2	A_2	t
V_i	1	2/3	1/2	1/3	1/6

This function for obtaining values for V_i entails the fact that V_i can never fall within the open interval (1, 2/3). If one wanted V_i values higher than 2/3 but less than unity, one would have to

move away from rank ordering and into some type of interval scale. This, of course, would further strain our notions of plausibility with regard to alternative hypotheses; it would be more difficult to make the required weighting assignments.

On the positive side, I think that this method, elementary as it is, would help us gain a sense of relative significance when we are trying to summarize several research studies on a given topic. Probably many people have a sense of this anyway—but one of the functions of logic is to explicate our intuitions.

V. Inductive Issues

The classical or Neyman-Pearson statistical theory does not, of course, assign probabilities to hypotheses. Thus, some might wonder whether the proposal offered here is aesthetically compatible with this theory. V_i should not be seen as a probability; rather it is a weight of _t_ against its alternatives. This kind of weighting is indigenous to classical statistics.

Both actual practice and the proposal of this paper fall nicely into Harman's view of induction as "inference to the best explanation": (1965).

> In making this inference one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis. In general, there will be several hypotheses which might explain the evidence, so one must be able to reject all such alternative hypothesis before one is warranted in making the inference. Thus, one infers, from the premise that a given hypothesis would provide a "better" explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true (p. 89).

I think that it is evident that Harman's inference to the best explanation is precisely the kind of inference at issue in the questions of internal validity. Moreover, we are considering the cases where we cannot "reject all such alternative hypotheses." Before we go to the public with claims of efficacy we will want to be able to reject all alternatives; but as we talk to each other we require some way of estimating the relative merits of alternative ways of viewing what happened in various studies. The account of estimating internal validity developed in this paper is proposed as such a way.

Finally, there is an interesting aspect of Harman's notion as it relates to an issue in the logic of science. Some philosophers believe that Harman's view is circular in that induction to the best explanation presupposes a way of determining "best" which is itself an inductive process. However, I will risk the following claim: while inference to the best explanation is shaky as a method of producing general scientific knowledge, it is both the method-in-use for determining internal validity and is a sound move for so doing when viewed logically. Within the confines of a single experiment, i.e., where there is no concern for generalizability, the only rationally defensible way of inferring "what happened" is the method described by Harman.

References

Gilbert H. Harman, The Inference to the Best Explanation, Philosophical Review 74 (1965), 88-95.

Nicholas Maxwell, The Rationality of Scientific Discovery, Part I (June) and II (Sept.). Philosophy of Science 41 (1974).